GeoDeepDive: A Cyberinfrastructure to Support Text and Data Mining

Ian Ross, Miron Livny, Tim Theisen
Center for High Throughput Computing
University of Wisconsin-Madison USA

Shanan E. Peters, John Czaplewski
Department of Geoscience
University of Wisconsin-Madison USA
How has biodiversity changed on Earth over geologic time?
Paleobiology Database

~10 continuous person years of effort
How has biodiversity changed on Earth over geologic time?

The "answer"

The combined effort has paid off!

Number of genera

<table>
<thead>
<tr>
<th>Cm</th>
<th>O</th>
<th>S</th>
<th>D</th>
<th>C</th>
<th>P</th>
<th>Tr</th>
<th>J</th>
<th>K</th>
<th>Pg</th>
<th>Ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>400</td>
<td>300</td>
<td>200</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time (Ma)

Citation indices

- Citations: All 17149, Since 2011 13608
- h-index: 62, 53
- i10-index: 367, 354

Official publications

- 2008: 5, 3
- 2009: 8, 6
- 2010: 10, 7
- 2011: 15, 10
- 2012: 20, 13
- 2013: 25, 16
- 2014: 30, 19
- 2015: 35, 22
- 2016: 40, 25
Can a machine reading system reproduce the PBDB?

- machine

  DeepDive v0.8.0
  Think about features, not algorithms.
  http://deepdive.stanford.edu

- humans

  The Paleobiology Database
  revealing the history of life
“Data entry” = feature identification and extraction

Journal Articles
“Data entry” = feature identification and extraction

… The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members...
“Data entry” = feature identification and extraction

… The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members ...

The Namurian Tsingyuan Formation from Ningxia
“Data entry” = feature identification and extraction

The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members...

<table>
<thead>
<tr>
<th>Entity1</th>
<th>Entity2</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Namurian</td>
<td>Tsingyuan Fm.</td>
<td>nn</td>
</tr>
<tr>
<td>Silesian</td>
<td>Tsingyuan Fm.</td>
<td>SameRow</td>
</tr>
</tbody>
</table>

Table: Age | Formation
---|---------
Silesian | Tsingyuan Formation
PaleoDeepDive

- From a collection of paleontological literature, extracts relations between biological taxa, geological formations, geographic locations, and geological time intervals

- "PaleoDeepDive performs comparably to humans in several complex data extraction and inference tasks and generates congruent synthetic results that describe the geological history of taxonomic diversity and genus-level rates of origination and extinction."

  - "A Machine Reading System for Assembling Synthetic Paleontological Database" (Peters, Zhang, Livny, Re)


- Also shows that the quality of the data inferences systematically improves as information is added.
We are here, but much of the data are over there.
Three major hurdles:

- Access to documents
- Processing power
- Dependability/repeatability
A shift in project ambitions...
Let’s build a “smart” library of TDM products!
Three Infrastructure Challenges

• Access to documents
  • Legal and responsible access to scientific literature

• Processing power
  • Need resources, automation, and flexibility
    • This framework should be useful for non-DeepDive applications as well!

• Dependability/Repeatability
  • Track the source of every word/sentence provided to an enduser, and always provide links back to the original content owner.
large-scale processing jobs
(using encrypted file system)

DeepDiveSubmit
TDM-ready library, supporting databases

controlled/authorized access

content owners/providers

WILEY
Publishers Since 1807

USGS
science for a changing world
ELSEVIER
CAMBRIDGE UNIVERSITY PRESS

controlled document fetching
(key/rate-based)

Secure@SWAMP
original PDF document storage
highly restricted access
large-scale **processing jobs**
(using encrypted file system)

- Website/API
- GitHub App
- DeepDiveSubmit
  TDM-ready library, supporting databases

- Database
- Secure@SWAMP
  TDM-ready library, supporting databases

controlled/authorized access

controlled document fetching (key/rate-based)

- content owners/providers
- Wiley Publishers Since 1807
- USGS science for a changing world
- Elsevier
- Cambridge University Press

Ian Ross — HTCondor Week 2016
Infrastructure Challenge 1: Legal and Responsible Access to Documents

• Working with the UW Library to draft and sign contracts with large publishers

• Strive to be “Good citizens”
  • Limit ourselves to an agreed-upon fetching rate
  • Providing feedback to publishers, as inconsistent data and system hiccups are discovered
  • Never provide the PDFs themselves to endusers! Only derived products and links back to the provider.
large-scale processing jobs
(using encrypted file system)

DeepDiveSubmit
TDM-ready library, supporting databases

Secure@SWAMP
original PDF document storage
highly restricted access

controlled/authorized access

controlled document fetching
(key/rate-based)

Website/API
GitHub App

TDM user
Ian Ross — HTCondor Week 2016

large-scale processing jobs
(using encrypted file system)

DeepDiveSubmit
encrypted PDFs for processing,
TDM-ready library, supporting databases

Secure@SWAMP
original PDF document storage
highly restricted access

controlled/authorized access

content owners/providers

controlled document fetching
(key/rate-based)

Website/API
GitHub App
DeepDiveSubmit
Database

TDM user
Infrastructure Challenge 2: Processing Power

- 1.2 million articles (+50,000 per week), 6 processing types
- Fairly small/short analysis jobs (3-5 minutes on average for OCR jobs, slightly longer for NLP)
- High throughput computing is exactly what we need!
- Use HTCondor and the UW CHTC resources for all of this processing work.

---

• The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members...

<table>
<thead>
<tr>
<th>Entity1</th>
<th>Entity2</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Namurian</td>
<td>Tsingyuan Fm.</td>
<td>nn</td>
</tr>
<tr>
<td>Silesian</td>
<td>Tsingyuan Fm.</td>
<td>SameRow</td>
</tr>
</tbody>
</table>
The Infrastructure Challenge — Processing

• Specific needs:

  • Automation and organization — 50,000 articles x 6 different processing types = Potential management nightmare
    • Database makes it easy to ID articles that need processing, keep track of products.

  • Security

  • Flexibility — New tools and document sources should be easy to add to the pipeline
    • New documents are easy — if there’s an entry in the database, they’ll get processed
The Infrastructure Challenge — Processing

• Specific needs:
  • Automation and organization: 50,000 articles x 6 different processing types = Potential management nightmare
  • Database makes it easy to identify articles that need processing and keep track of products.
  • Security
  • Flexibility — New tools and document sources should be easy to add to the pipeline
  • New documents are easy — if there's an entry in the database, they'll get processed

• Provided by HTCondor!
  • DAGMan w/ postscript to help stay organized (rescue files, dag-level-throttling)
  • Encrypted filesystem ensures PDFs won’t be left exposed on the execute nodes
Throughput Statistics

- With a fetch rate of 50,000 articles/week and 6 current processing types, “steady state” requires ~5000 cpu hours per day

- Also have the capability (and resources!) to deploy new processing types across the entire corpus
  - Recently deployed a new version of the Stanford CoreNLP tool to all documents (1 million at the time)
    - Took ~3 weeks to process 1 million documents, while still staying up to date with the other processing types
large-scale processing jobs
(using encrypted file system)

DeepDiveSubmit
TDM-ready library, supporting databases

Website/API
GitHub App

encrypted PDFs for processing,

controlled/authorized access

Secure@SWAMP
original PDF document storage
highly restricted access

content owners/providers

controlled document fetching
(key/rate-based)
large-scale processing jobs
(using encrypted file system)

Website/API
GitHub App
DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

content owners/providers
WILEY
Publishers Since 1807
USGS
science for a changing world
ELSEVIER
CAMBRIDGE UNIVERSITY PRESS

controlled document fetching
(key/rate-based)

controlled/authorized access

Ian Ross — HTCondor Week 2016
What can be done with all this data?

• PaleoDeepDive showed that machine reading can infer facts and build a database comparable to years of human effort

• Even without bringing machine learning into the picture, there’s a huge amount of value in the sentence-level data and NLP products!
e.g. Space-time index of the literature
https://macrostrat.org/burwell/
Tunnel City Group, Elk Mound Group

**Age:** Dresbachian - Trempealeauan - (499.95 - 487.175 Ma)

**Thickness:** 0 - 150 m

**PBDB Collections:** 4

**Unit IDs:** 6052, 6053, 6104, 6105, 6106, 6117, 6119, 6120, 6126, 6127, 6128, 6131, 6132, 6133, 6141, 6142, 6143, 6144, 6146, 6252

**Reference:** Macrostrat.org

---

**Professional Paper** USGS

Young, H. L., Siegel, D. I., 1992. Hydrogeology of the Cambrian-Ordovician aquifer system in the northern Midwest, United States, with a section on ground-water quality.

---

...underlying Eau Claire Formation... and its partial equivalent to the southwest, the Bonneterre Formation...

...Siltstone and shale are fairly common in the upper part of the Eau Claire Formation but less so in its...

...the Eau Claire Formation in northern Illinois. The aquifer increases greatly in thickness and the...

...Biogenic Depositional shelf St. Lawrence Formation Tunnel City Group Van Oser Member Norwalk...

Bridge, Josiah, 1937. The correlation of the Upper Cambrian sections of Missouri and Texas with the section in the upper Mississippi Valley.
e.g. New synthetic results
GeoDeepDive + Macrostrat
tuple extraction: lots of entities, NLP features link them
Stromatolite prevalence in the geologic record

normalized to marine sediments

stratigraphic abundance

age (Ma)

Parc  Marc  Narc  Pptz  Mptz  Nptz  Pz  Mz

3500  3000  2500  2000  1500  1000  500  0
Conclusions — Key Infrastructure Features

• Automated document fetching at arbitrary maximum rates determined by content providers (e.g., Elsevier 10K/week/API key)

• Secure document storage; encrypted processing methods to protect content owners/providers

• HTC infrastructure to run core tools (e.g., NLP, OCR, table recognition/parsing, image analysis), with flexibility and power to add more.

• API layer with basic capacity to identify documents of potential relevance to a project, with initial results returned as (augmented) bibJSON

• Packaging and delivery of analysis-ready sentence data (e.g., PostgreSQL database of NLP results); everything traceable back to specific sources (original URL and locations within documents)
Lots of ways to get involved!

• Identify and help retrieve documents from content owners (e.g., museum publications series, society publications, open-access content)

• Write TDM applications that can facilitate your work/science and do cool things; *we will help you!*

• Develop tools for parsing/reading documents in your area of work; develop comprehensive dictionaries of terms in your field and make them accessible to us so we can pre-index the literature

• Leverage our APIs in your applications (just let us know, we might help)!
Questions?

• http://www.geodeepdive.org

• iross@cs.wisc.edu